Linkspotter : an R package for correlations analysis and visualization

Alassane Samba

Linkspotter : outil interactif d'exploration et de visualisation de corrélations

Alassane Samba

Orange Labs
2 avenue Pierre Marzin, 22300 Lannion, France
alassane.samba@orange.com

Mots clefs : Corrélation, Information mutuelle, Graphe dynamique, Visualisation, Interface utilisateur.

L'exploration des relations entre les variables est une étape importante dans le processus d'exploration de données. Il permet au Data scientist de mieux comprendre les phénomènes décrits dans les jeux de données. Cela aide également à éviter le sur-apprentissage et à se prémunir contre le « concept drift » dans la modélisation.

Afin de faciliter l'exploration des données, Linkspotter est un package R offrant plusieurs fonctionnalités permettant d'analyser et de visualiser de manière exhaustive en utilisant un graphe toutes les corrélations bi-variées d'un fichier de données.

Ses fonctionnalités principales sont:

- le calcul de plusieurs matrices de corrélation correspondant à différents coefficients
- le partitionnement des variables par apprentissage non supervisé.

Il offre également une interface utilisateur complète, conviviale et personnalisable permettant de :

- visualiser les corrélations à l'aide d'un graphe (les variables correspondant aux noeuds et les corrélations correspondant aux arcs). C'est une nouvelle approche de visualisation qui est proposée plus conviviale que l'affichage d'une matrice de corrélations coloriée.
- visualiser la distribution de chaque variable grâce à son histogramme ou à son diagramme en barres.
- visualiser un lien entre un couple de variables à l'aide de nuage de points, de boîtes-à-moustaches, etc.

En outre, un nouveau coefficient de corrélation que nous appelons Maximal Normalized Mutual Information (MaxNMI) basé sur une discrétisation supervisée Equal-Freq des variables continues est également introduit par Linkspotter. L'intérêt de ce coefficient est qu'il peut être calculé et comparé quel que soit le type de couple de variables (continue vs catégorielle, continue vs continue, catégorielle vs catégorielle).

Tout en offrant de nouveaux algorithmes et méthodes, Linkspotter utilise et combine harmonieusement des fonctionnalités provenant d'autres packages R, à savoir infotheo, minerva, energy, mclust, shiny, visNetwork et rAmCharts.

Le code [1] et une démonstration [2] de Linkspotter sont disponibles en ligne.

Références

[1] https://github.com/sambaala/linkspotter
[2] http://linkspotter.sigmant.net